10/10/00 JC03 U.S. PTO

JC826 U.S. PTO 09/684331 10/10/00

LAW OFFICES
SUGHRUE, MION, ZINN, MACPEAK & SEAS, PLLC
2100 PENNSYLVANIA AVENUE, N.W.
WASHINGTON, DC 20037-3213
TELEPHONE (202) 293-7060
FACSIMILE (202) 293-7860
www.sughrue.com

October10, 2000

BOX PATENT APPLICATION
Assistant Commissioner for Patents
Washington, D.C. 20231

Re: Application of Katsumi AMANO, Shisei CHO, Soichi TOYAMA and Hiroyuki ISHIHARA
SPEECH SYNTHESIS METHOD
Our Ref. Q61111

Dear Sir:

Attached hereto is the application identified above including sixteen (16) sheets of the specification, claims and six (6) sheets of informal drawings. **The requisite U.S. Government Filing Fee, executed Declaration and Power of Attorney and Assignment will be submitted at a later date.**

The Government filing fee is calculated as follows:

| | | | | | | |
|---|---|---|---|---|---|---|
| Total claims | 4 - 20 | = | | x | $18.00 | = $.00 |
| Independent claims | 1 - 3 | = | | x | $80.00 | = $.00 |
| Base Fee | | | | | | $710.00 |
| **TOTAL FEE** | | | | | | **$710.00** |

Priority is claimed from October 15, 1999 based on JP Application No. 11-294357. The priority document will be filed at a later date.

Respectfully submitted,
SUGHRUE, MION, ZINN,
MACPEAK & SEAS, PLLC
Attorneys for Applicant

By: _____
Darryl Mexic
Registration No. 23,063

DM:clf

# SPEECH SYNTHESIS METHOD

## 1. FIELD OF THE INVENTION

The present invention relates to a speech synthesis method for artificially generating speech waveform signals.

## 2. BACKGROUND OF THE RELATED ART

Speech waveforms of natural speech can be expressed by connecting basic units which are made by continuously connecting phonemes, one vowel (V) and one consonant (C) in a form such as "CV", "CVC" or "VCV".

Accordingly, a conversation can be created by means of synthetic speech by processing and registering such phonemes as data (phoneme data) in advance, reading out phoneme data corresponding to a conversation from the registered phoneme data in sequence, and generating sounds corresponding to respective read-out phoneme data.

To create a database based on the above-mentioned phoneme data, firstly, a given document is read by a person, and his/her speech is recorded. Then, speech signals reproduced from the recorded speech are divided into the above-mentioned phonemes. Various data indicative of these phonemes are registered as phoneme data. Then, in order to synthesize the speech, respective speech data is connected and supplied as a serial speech.

However, respective connected phonemes are segmented from the separately recorded speeches. Hence, irregularities exist in the vocal power with which the phonemes are uttered.

Therefore, a problem arises that synthesized speech is unnatural when the uttered phonemes are merely connected together.

An object of the present invention is to provide a speech synthesizing method for generating natural sounding synthetic speech.

## SUMMARY OF THE INVENTION

It is an object of the present invention to provide a method for synthesizing speech with an apparatus comprising a sound source for generating a frequency signal, a vocal tract filter for generating speech waveform signals by filtering the frequency signal with filter characteristics corresponding to a linear predictive coefficient based on respective phonemes.

In one aspect of the invention, a method comprises the steps of: dividing said phonemes into a plurality of frames having a predetermined time length, summing squares of speech samples in one of said plurality of frames for each frame as a frame power value, standardizing frame power values at head and tail frames in one phoneme to predetermined values, respectively, to obtain a frame power value of an n-th frame, summing squares of signal levels of a frame in said frequency signal to obtain a frame power correction value, providing a speech envelope signal by means of a function having variables of said standardized frame power values and said frame power correction value, and adjusting an amplitude level of said speech waveform signal as a function of the speech envelope

signal.

As described above, the levels of the head and tail portions of respective phonemes are always maintained at predetermined levels without substantially deforming the synthesized speech waveform. Therefore, phonemes are connected together smoothly so that natural sounding synthesized speeches can be generated.

BRIEF DESCRIPTION OF THE DRAWINGS

The aforementioned aspects and other features of the invention are explained in the following description, taken in connection with the accompanying drawing figures wherein:

Fig. 1 is a block diagram showing a speech synthesis apparatus according to the present invention,

Fig. 2 is a block diagram showing an apparatus for generating phoneme data and speech synthesis parameters,

Fig. 3 is a flow chart showing steps for generating phoneme data,

Fig. 4 is a view showing a memory map in a memory 33,

Fig. 5 is a flow chart showing steps for calculating speech synthesis parameters, and

Fig. 6 is a view showing a speech synthesis control routine based on a speech synthesis method of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

Fig. 1 is a block diagram showing a text speech

synthesis device for reading a given document (text) by synthesizing the speech by means of a method according to the present invention.

In Fig. 1, a text analyzing circuit 21 generates intermediate language character string information including information such as accents and phrases peculiar to respective languages in a character string based on inputted text signals. The text analyzing circuit 21 then supplies intermediate language character string signals CL corresponding to the above information to a speech synthesis control circuit 22.

A phoneme data memory 20, a RAM (Random Access Memory) 27, and a ROM (Read Only Memory) 28 are connected to the speech synthesis control circuit 22.

The phoneme data memory 20 stores phoneme data corresponding to various phonemes which have been sampled from actual human voice, and speech synthesizing parameters (standardized frame power values and mean frame power values) used for the speech synthesis.

A sound source module 23 is provided with a pulse generator 231 for generating impulse signals having a frequency corresponding to a pitch frequency designating signal K supplied from the speech synthesis control circuit 22, and a noise generator 232 for generating noise signals carrying an unvoiced sound. The sound source module 23 alternatively selects the impulse signal and the noise signal in response to a sound source selection signal $S_V$ supplied from the speech synthesis control circuit 22. The sound source

module 23 then supplies the selected signal as a frequency signal Q to a vocal tract filter 24.

The vocal tract filter 24 may include a FIR (Finite Impulse Response) digital filter, for example. The vocal tract filter 24 filters a frequency signal Q supplied from the sound source module 23 with a filtering coefficient corresponding to a linear predictive code signal LP supplied from the speech synthesis control circuit 22, thereby generating a speech waveform signal $V_F$.

An amplitude adjustment circuit 25 generates an amplitude adjustment waveform signal $V_{AUD}$ by adjusting the amplitude of a speech waveform signal $V_F$ to a level based on a speech envelope signal $V_m$ supplied from the speech synthesis control circuit 22. The amplitude adjustment circuit 25 then supplies the amplitude adjustment waveform signal $V_{AUD}$ to a speaker 26. The speaker 26 generates an acoustic output corresponding to the amplitude adjustment waveform signal $V_{AUD}$. That is, the speaker 26 generates the reading speeches based on the input text signals as explained hereinafter.

A method will be described hereinafter for generating the above-mentioned phoneme data and speech synthesis parameters stored in the phoneme data memory 20.

Fig. 2 is a block diagram showing an apparatus for generating speech synthesis parameters.

In Fig. 2, a speech recorder 32 records a human speech received by a microphone 31. The speech recorder 32 supplies speech signals reproduced from the recorded speech to a

phoneme data generating device 30.

The phoneme data generating device 30 sequentially samples a speech signal supplied from the speech recorder 32 to generate a speech sample. The phoneme data generating device 30 then stores the signals in a predetermined domain in a memory 33. The phoneme data generating device 30 then executes steps for generating phonemes, as shown in Fig. 3.

In Fig. 3, the phoneme data generating device 30 reads out speech samples stored in the memory 33 in sequence. The phoneme data generating device 30 then divides the series of speech samples into phonemes such as "VCV" (step S1).

For example, a Japanese spoken phrase "mokutekichi ni" is segmented to mo/oku/ute/eki/iti/ini/i. The Japanese spoken phrase "moyosimono" is segmented to mo/oyo/osi/imo/ono/ono/o. The Japanese spoken phrase "moyorino" is segmented to mo/oyo/ori/ino/o. The Japanese spoken phrase "mokuhyono" is segmented to mo/oku/uhyo/ono/o.

Subsequently, the phoneme data generating device 30 divides each segmented phoneme into frames of a predetermined length, for example, 10ms (step S2). Control information including a name of the phoneme to which each frame belongs, a frame length of the phoneme, and the frame number is added to each divided frame. The above frame is then stored in a given domain of the memory 33 (step S3). Then, the phoneme data generating device 30 analyzes a linear predictive coding LPC on every frame with respect to the waveform of each phoneme to generate a linear predictive coding coefficient (hereinafter

called "LPC coefficient") of 15 orders. The resultant coefficient is stored in a memory domain 1 of the memory 33 as shown in Fig. 4 (step S4). It should be noted that the resultant LPC coefficient in step S4 is a so-called speech spectral envelope parameter corresponding to a filter coefficient of the vocal tract filter 24. Subsequently, the phoneme data generating device 30 reads out the LPC coefficient in the memory domain 1 of the memory 33, and supplies the LPC coefficient as the phoneme data (step S5). This phoneme data is stored in the phoneme data memory 20.

Then, the phoneme data generating device 30 calculates speech synthesis parameters as shown in Fig. 5 on respective phonemes stored in the memory 33.

In Fig. 5, the phoneme data generating device 30 calculates the sum of all squares of speech sample values in each frame in one phoneme that is subject to processing (hereinafter called "subject phoneme") in order to generate a speech power of the frame. Then, as shown in Fig. 4, the speech power is stored in a memory domain 2 of the memory 33 as a frame power PC (step S12).

Subsequently, the phoneme data generating device 30 stores "0" indicative of the head frame number in a built-in register n (not shown) (step S13). Then, the phoneme data generating device 30 generates the relative position in the subject phoneme of the frame n indicated by the frame number stored in the built-in register n (step S14). The relative position is expressed by the following formula:

$$r = (n - 1)/N$$

wherein, r: relative position, and

N: the number of all frames in the subject phoneme.

Then, the phoneme data generating device 30 reads out the frame power PC in the frame n from the memory domain 2 of the memory 33 shown in Fig. 4 (step S15). The phoneme data generating device 30 reads out the frame powers corresponding to the head and tail frames of the subject phoneme as the head and tail frame powers $P_a$ and $P_b$, respectively, among the frame powers $P_c$ in the memory domain 2 (step S16).

Then, the phoneme data generating device 30 generates a standardized frame power $P_n$ in the frame n indicated by a built-in register n, by executing the following calculation (1) using the head and tail frame powers Pa, Pb, the frame power Pc obtained in step S15 and the relative position r.

$$P_n = P_c / [(1 - r) P_a + r P_b] \qquad (1)$$

Then, the phoneme data generating device 30 stores the standardized frame power Pn in a memory domain 3 of the memory 33 (step S17).

That is, the phoneme data generating device 30 generates the frame power value in the frame n when the frame power $P_c$ in the tail frame of this subject phoneme is set to "1".

Then, the phoneme data generating device 30 reads out the LPC coefficient corresponding to the frame n indicated by the built-in register n from the memory domain 1 of the memory 33 shown in Fig. 4. The phoneme data generating device 30 then

generates power frequency characteristics in the frame n based on the LPC coefficient (step S18). Thereafter, the phoneme data generating device 30 samples a power value from the power frequency characteristics every predetermined frequency interval, and then stores the average value of these power values as a mean frame power $G_f$ in a memory domain 4 of the memory 33 shown in Fig. 4 (step S19).

Then, the phoneme data generating device 30 adds "1" to the frame number n stored in the built-in register n to generate a new frame number n, the new frame number n replacing the previous frame number n, and stores the new frame number n in the built-in register n by substitution (step S20). Subsequently, the phoneme data generating device 30 determines whether the frame number stored in the built-in register n equals (N-1) (step S21).

In step S21, if the frame number stored in the built-in register n does not equal (N-1), the phoneme data generating device 30 returns to the step S14, and repeats the above-mentioned operation. Such an operation stores the standardized frame power $P_n$ and the mean frame power $G_f$ corresponding to each of the head frame to $(N-1)^{th}$ frames of a subject phoneme in the memory domains 3 and 4, as shown in Fig. 4.

In the step S21, if the frame number stored in the built-in register n equals (N-1), the phoneme data generating device 30 respectively reads out the standardized frame power $P_n$ and the mean frame power $G_f$ stored in the memory domains 3 and 4 of the memory 33 shown in Fig. 4, and outputs the

standardized frame power $P_n$ and the mean frame power $G_f$ (step S23). The standardized frame power $P_n$ and the mean frame power $G_f$ are stored in the phoneme data memory 20 as speech synthesis parameters.

That is, the respective phoneme data obtained by the procedure shown in Fig. 3 is associated with the standardized frame power $P_n$ and the mean frame power $G_f$ obtained by the procedure shown in Fig. 5 to store the resultant data in the phoneme data memory 20.

The speech synthesis control circuit 22 shown in Fig. 1 receives the phoneme data and speech synthesis parameters corresponding to the intermediate language characters string signals CL from the text analyzing circuit 21, by using software stored in the ROM 28. The speech synthesis control circuit 22 then controls speech synthesis as explained hereinafter.

The speech synthesis control circuit 22 divides segments of the intermediate language characters string signals CL into phonemes consisting of "VCV", and then receives the phoneme data corresponding to respective phonemes from the phoneme data memory 20 sequentially. The speech synthesis control circuit 22 then supplies a pitch frequency designation signal K for designating the pitch frequency to the sound source module 23. Then, the speech synthesis control circuit 22 synthesizes the speech on respective phoneme data in order of the reading from the phoneme data memory 20.

Fig. 6 shows a speech synthesizing control procedure.

In Fig. 6, the speech synthesis control circuit 22 selects the data for one phoneme subject to be processed (hereinafter called "subject phoneme data) in the received order as mentioned above. The speech synthesis control circuit 22 then stores "0" indicative of the head frame number in the phoneme data in the built-in register n (not shown) (step S101). Subsequently, the speech synthesis control circuit 22 supplies a sound source selection signal $S_v$ to the sound source module 23 (step S102). The sound source selection signal $S_v$ indicates whether the phoneme corresponding to the above-mentioned subject phoneme data is a voiced sound or an unvoiced sound. Depending on the sound source selection signal $S_v$, the sound module 23 generates as a frequency signal Q one of a noise signal and an impulse signal having a frequency designated by the pitch frequency designation signal K.

Subsequently, the speech synthesis control circuit 22 samples the frequency signal Q supplied from the sound source module 23 for every predetermined interval. The control circuit 22 then calculates the sum of squares of respective sample values in a frame to generate a frame power correction value $G_s$. Then, the speech synthesis control circuit 22 stores the frame power correction value $G_s$ in a built-in register G (not shown) (step S103). Then, the speech synthesis control circuit 22 supplies the LPC coefficient to the vocal tract filter 24 as the linear predictive coding signal LP (step S104). It is noted that the LPC coefficient corresponds to the frame n indicated by the built-in register n in the subject

phoneme data. Then, the speech synthesis control circuit 22 reads out the standardized frame power $P_n$ and the mean frame power $G_f$ corresponding to the frame n indicated by the above-mentioned built-in register n in the subject phoneme data from the phoneme data memory 20 (step S105). Thereafter, the speech synthesis control circuit 22 calculates a speech envelope signal $V_m$, by the following computation with the standardized frame power $P_n$, the mean frame power $G_f$, and the frame power correction value Gs stored in the built-in register G. The speech synthesis control circuit 22 then supplies the speech envelope signal $V_m$ to an amplitude adjustment circuit 25 (step S106).

$$V_m = \sqrt{P_n / (G_s G_f)}$$

By means of the step S106, the amplitude adjustment circuit 25 adjusts the amplitude of the speech waveform signal $V_f$ supplied from the vocal tract filter 24 to a level corresponding to the above-mentioned speech envelope signal $V_m$. Since the connecting portions of respective phonemes are always maintained at a predetermined level through this amplitude adjustment, the connection of phonemes becomes smooth and hence, natural sounding synthesized speech is produced.

Subsequently, the speech synthesis control circuit 22 determines whether the frame number n stored in the built-in register n is smaller than the total number of frames in the subject phoneme data N by 1, that is, whether the frame number

n equals (N - 1) (step S107). In the step S107, if it is determined that n does not equal (N-1), the speech synthesis control circuit 22 adds "1" to the frame number stored in the built-in register n, and stores this value as a new frame number in the built-in register n by substitution (step S108). After the step S108, the speech synthesis control circuit 22 returns to the step S103, and then repeats the above-mentioned operation.

On the other hand, in step S107, if it is determined that the frame number n stored in the built-in register n does not equal (N-1), the speech synthesis control circuit 22 returns to the step S101, and repeats the phonemic synthesis process to next phoneme data in the same manner.

The present invention has been explained heretofore in conjunction with the preferred embodiment. However, it should be understood that those skilled in the art could easily conceive various other embodiments and modifications and that such embodiments and modifications fall within the scope of the appended claims.

WHAT IS CLAIMED IS

1. A method for synthesizing speech with an apparatus comprising a sound source for generating a frequency signal, a vocal tract filter for filtering said frequency signal to generate a speech waveform signal, said filter having characteristics corresponding to a linear predictive coefficient calculated from respective phonemes in a phoneme series, comprising the steps of:

dividing said phonemes into a plurality of frames having a predetermined time length,

summing squares of speech samples in one of said plurality of frames for each frame as a frame power value,

standardizing frame power values at head and tail frames in one phoneme to predetermined values, respectively, to obtain a frame power value of an n-th frame,

summing squares of signal levels of a frame in said frequency signal to obtain a frame power correction value, and

providing a speech envelope signal by means of a function having variables of said standardized frame power values and said frame power correction value, and adjusting an amplitude level of said speech waveform signal as a function of the speech envelope signal.

2. A method according to claim 1, further comprising:

providing power frequency characteristics based on said linear predictive coefficient corresponding to said n-th frame,

calculating an average value of power values sampled from said power frequency characteristics at a predetermined

frequency interval as a mean frame power value,

calculating a speech waveform signal by means of a function having variables of said standardized frame power value, said frame power correction value and said mean frame power value, and

adjusting an amplitude of said speech waveform signal as a function of said speech envelope signal.

3. A method according to claim 2, wherein said function is expressed;

$$V_m = \sqrt{P_n / (G_s G_f)}$$

wherein $P_n$ is said standardized frame power value, $G_s$ is said frame power correction value, and $G_f$ is said mean frame power value.

4. A method according to claim 1, wherein said frequency signal includes an impulse signal carrying a voiced sound and a noise signal carrying an unvoiced sound.

ABSTRACT

A speech synthesizing method which synthesizes speech naturally is disclosed. Standardized frame power values of an n-th frame is calculated when frame power values at head and tail frames in a phoneme are standardized. An average value of the power values sampled from the power frequency characteristics in the n-th frame at a predetermined frequency interval is set as a mean frame power value. A sum of squares of signal levels in one frame of a frequency signal from a sound source is calculated as a frame power correction value. A speech envelope signal is calculated as a function having variables of the standardized frame power values, the frame power correction value and the mean frame power value. The speech envelope signal adjusts the amplitude level of a speech waveform signal supplied from a vocal tract filter according to the level of the speech envelope signal.

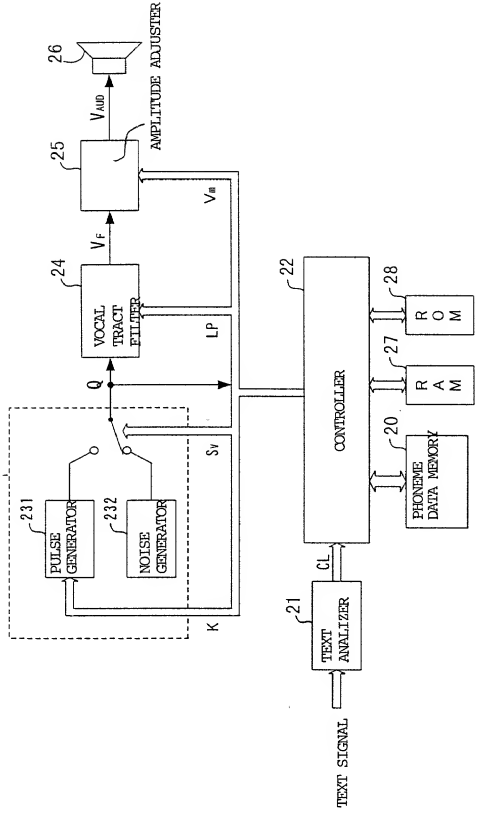
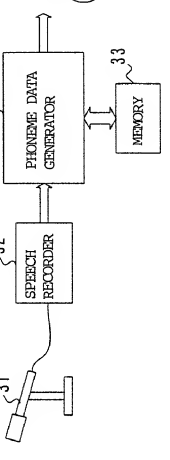

TEXT SIGNAL
TEXT ANALYZER
21
CL
CONTROLLER
22
PHONEME DATA MEMORY
20
RAM
27
ROM
28
23
PULSE GENERATOR
231
NOISE GENERATOR
232
K
Sv
Q
LP
VOCAL TRACT FILTER
24
VF
Va
25
AMPLITUDE ADJUSTER
VAUD
26

FIG. 1

FIG. 2


31
32
SPEECH RECORDER
30
PHONEME DATA GENERATOR
33
MEMORY
PHONEME DATA
SPEECH SYNTHESIS PARAMETER
STANDARDIZED FRAME POWER VALUE
MEAN FRAME POWER VALUE

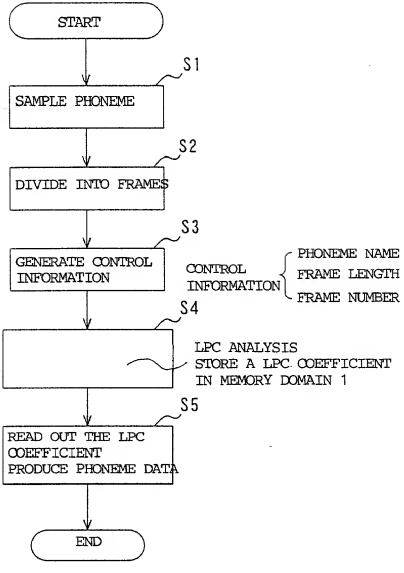
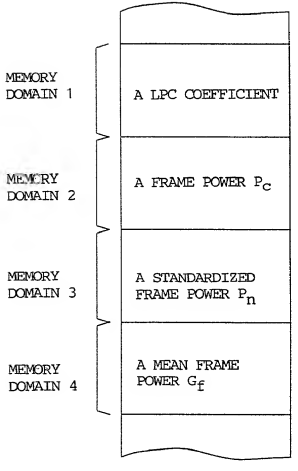

START
S1
SAMPLE PHONEME
S2
DIVIDE INTO FRAMES
S3
GENERATE CONTROL INFORMATION
CONTROL INFORMATION
PHONEME NAME
FRAME LENGTH
FRAME NUMBER
S4
LPC ANALYSIS
STORE A LPC COEFFICIENT IN MEMORY DOMAIN 1
S5
READ OUT THE LPC COEFFICIENT
PRODUCE PHONEME DATA
END

FIG. 3

# FIG. 4


MEMORY DOMAIN 1
A LPC COEFFICIENT
MEMORY DOMAIN 2
A FRAME POWER $P_c$
MEMORY DOMAIN 3
A STANDARDIZED FRAME POWER $P_n$
MEMORY DOMAIN 4
A MEAN FRAME POWER $G_f$

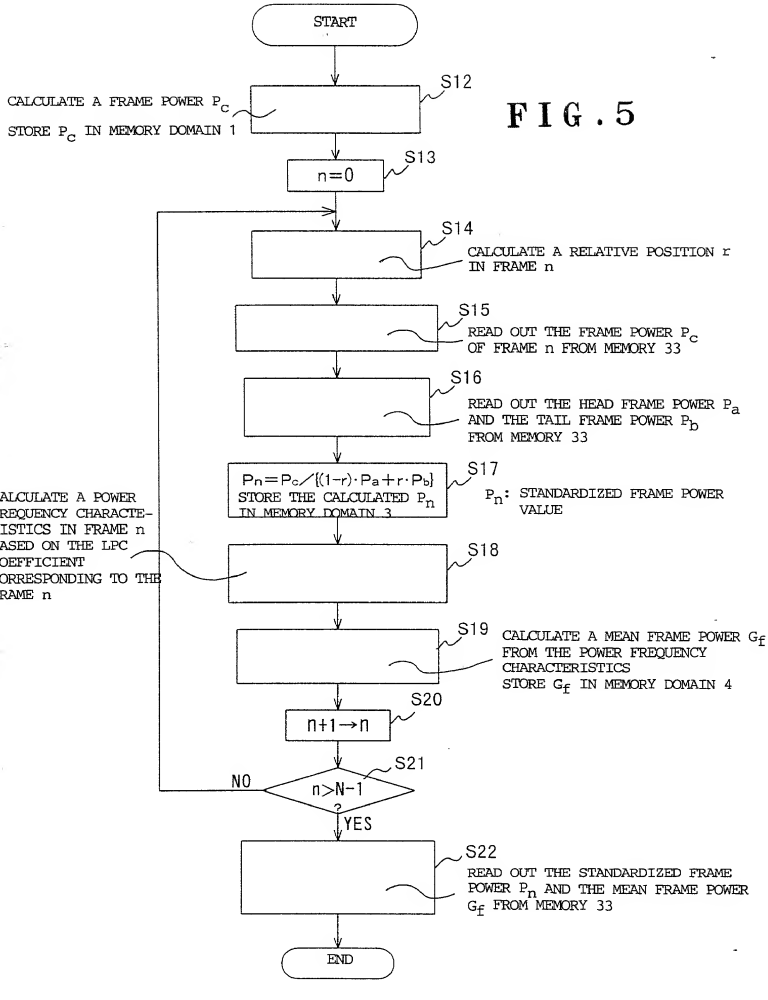
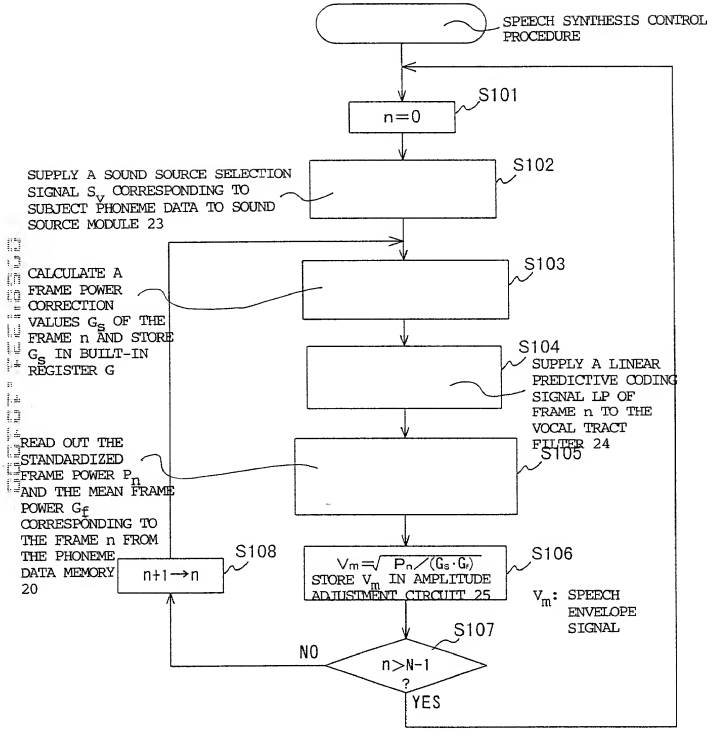

START
S12
CALCULATE A FRAME POWER $P_c$
STORE $P_c$ IN MEMORY DOMAIN 1
S13
n=0
S14
CALCULATE A RELATIVE POSITION r IN FRAME n
S15
READ OUT THE FRAME POWER $P_c$ OF FRAME n FROM MEMORY 33
S16
READ OUT THE HEAD FRAME POWER $P_a$ AND THE TAIL FRAME POWER $P_b$ FROM MEMORY 33
$P_n=P_c/\{(1-r)\cdot P_a+r\cdot P_b\}$
STORE THE CALCULATED $P_n$ IN MEMORY DOMAIN 3
S17
$P_n$: STANDARDIZED FRAME POWER VALUE
CALCULATE A POWER FREQUENCY CHARACTERISTICS IN FRAME n BASED ON THE LPC COEFFICIENT CORRESPONDING TO THE FRAME n
S18
S19
CALCULATE A MEAN FRAME POWER $G_f$ FROM THE POWER FREQUENCY CHARACTERISTICS
STORE $G_f$ IN MEMORY DOMAIN 4
S20
n+1→n
S21
NO
n>N-1 ?
YES
S22
READ OUT THE STANDARDIZED FRAME POWER $P_n$ AND THE MEAN FRAME POWER $G_f$ FROM MEMORY 33
END

FIG. 5

# FIG. 6


SPEECH SYNTHESIS CONTROL PROCEDURE
S101
n=0
S102
SUPPLY A SOUND SOURCE SELECTION SIGNAL $S_V$ CORRESPONDING TO SUBJECT PHONEME DATA TO SOUND SOURCE MODULE 23
S103
CALCULATE A FRAME POWER CORRECTION VALUES $G_s$ OF THE FRAME n AND STORE $G_s$ IN BUILT-IN REGISTER G
S104
SUPPLY A LINEAR PREDICTIVE CODING SIGNAL LP OF FRAME n TO THE VOCAL TRACT FILTER 24
S105
READ OUT THE STANDARDIZED FRAME POWER $P_n$ AND THE MEAN FRAME POWER $G_f$ CORRESPONDING TO THE FRAME n FROM THE PHONEME DATA MEMORY 20
S106
$V_m=\sqrt{P_n/(G_s \cdot G_f)}$
STORE $V_m$ IN AMPLITUDE ADJUSTMENT CIRCUIT 25
$V_m$: SPEECH ENVELOPE SIGNAL
S107
n>N-1 ?
NO
YES
S108
n+1→n